A Probabilistic Genre-independent Model of Pronominalization

A Probabilistic Genre-Independent Model of Pronominalization

Michael Strube
European Media Laboratory GmbH
Villa Bosch
Schloß-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
Michael.Strube@eml.villa-bosch.de

Maria Wolters
Inst. f. Kommunikationsforschung u. Phonetik
Universität Bonn
Poppelsdorfer Allee 47
53115 Bonn, Germany
wolters@ikp.uni-bonn.de

Abstract

Our aim in this paper is to identify genre-independent factors that influence the decision to pronominalize. Results based on the annotation of twelve texts from four genres show that only a few factors have a *strong* influence on pronominalization across genres, i.e. distance from last mention, agreement, and form of the antecedent. Finally, we describe a probabilistic model of pronominalization derived from our data.

1 Introduction

Generating adequate referring expressions is an active research topic in Natural Language Generation. Adequate referring expressions are those that enable the user to quickly and unambiguously identify the discourse entity that the expression co-specifies with. In this paper, we concentrate on an important aspect of that question, which has received less attention than the question of anaphora resolution in discourse interpretation, i.e., when is it feasible to pronominalize?

Our aim is to identify the central factors that influence pronominalization across genres. Section 2 motivates and presents the factors that were investigated in this study: distance from last mention, parallelism, ambiguity, syntactic function, agreement, sortal class, syntactic function of the antecedent and form of the antecedent. Our analyses are based on a corpus of twelve texts from four different genres with a total of more than 24,000 words and 7126 referring expressions (Section 3). The results of the statistical analyses are summarized in Section 4. There are strong statistical associations between each of the factors and pronominalization. Only when we combine them into a probabilistic model we can identify those factors whose contribution is really important, i.e. distance from last mention, agreement, and to a certain degree form of the antecedent. Since these factors can be annotated relatively cheaply, we conclude that it is possible to develop reasonable statistical pronominalization algorithms.

2 Factors in Pronoun Generation

2.1 Previous Work

Lately, a number of researchers have done corpus-based work on NP generation and pronoun resolution, and a number of studies have found differences in the frequency of both personal and demonstrative pronouns across genres. However, none of these studies compares the influence of different factors on pronoun *generation* across genres.

Recently, Poesio et al. (1999) have described a corpus-based approach to statistical NP generation. While they ask the same question as previous researchers (e.g. Dale (1992)), their methods differ from traditional work on NP generation. Poesio et al. (1999) use two kinds of factors: (1) factors related to the NP under consideration such as agreement information, semantic factors, and discourse factors, and (2) factors related to the antecedent, such as animacy, clause type, thematic role, proximity, etc. Poesio et al. (1999) report that they were not able to annotate many of these factors reliably. On the basis of these annotations, they constructed decision trees for predicting surface forms of referring expressions based on these factors - with good results: all 28 personal pronouns in their corpus were generated correctly. Unfortunately, they do not evaluate the contribution of each of these factors, so we do not know which ones are important.

Work on corpus-based approaches to anaphora resolution is more numerous. Ge et al. (1998) describe a supervised probabilistic pronoun resolution algorithm which is based on complete syntactic information. The factors they use include distance from last mention, syntactic function and context, agreement information, animacy of the referent, a simplified notion of selectional restrictions,

Agree	Agreement in person, gender, and number
Syn	Syntactic function
Class	Sortal Class (cf. Tab. 2)
SynAnte	Syntactic function of antecedent. "F" for first mention, "N" for deadend
FormAnte	Form of antecedent (pers. pron., poss. pron., def. NP, indef. NP, proper name)
Dist	Distance to last mention in units
Dist4	Dist reduced to 4 values (deadend, Dist=0, Dist=1, Dist>=2)
Par	Parallelism (Syn=SynAnte)
Ambig	Number of competing discourse entities

Table 1: Overview of factors

and the length of the coreference chain. Cardie & Wagstaff (1999) describe an unsupervised algorithm for noun phrase coreference resolution. Their factors are taken from Ge et al. (1998), with two exceptions. First, they replace complete syntactic information with information about NP bracketing. Second, they use the sortal class of the referent which they determine on the basis of WordNet (Fellbaum, 1998).

There has been no comparison between corpus-based approaches for anaphora resolution and more traditional algorithms based on focusing (Sidner, 1983) or centering (Grosz et al., 1995) except for Azzam et al. (1998). However, their comparison is flawed by evaluating a syntax-based focus algorithm on the basis of insufficient syntactic information. For pronoun generation, the original centering model (Grosz et al., 1995) provides a rule which is supposed to decide whether a referring expression has to be realized as a pronoun. However, this rule applies only to the referring expression which is the backward-looking center (*Cb*) of the current utterance. With respect to all other referring expression in this utterance centering is underspecified.

Yeh & Mellish (1997) propose a set of handcrafted rules for the generation of anaphora (zero and personal pronouns, full NPs) in Chinese. However, the factors which appear to be important in their evaluation are similar to factors described by authors mentioned above: distance, syntactic constraints on zero pronouns, discourse structure, salience and animacy of discourse entities.

2.2 Our Factors

The factors we investigate in this paper only rely on annotations of NPs and their co-specification relations. We did not add any discourse structural annotation, because (1) the texts are extracts from larger texts which are not available to us, and (2) we have not yet found a labelling scheme for discourse structure that has an inter-coder reliability comparable to the MUC coreference annotation scheme.

Based on our review of the literature and relevant work in linguistics (for sortal class, mainly Fraurud (1996) and Fellbaum (1998)), we have chosen the nine factors listed in Table 1. Methodologically, we distinguish two kinds of factors:

NP-level factors are independent from co-specification relations. They depend on the semantics of the discourse entity or on discourse information supplied for the NP generation algorithm by the NLG system. Typical examples are NP agreement by gender, number, person and case, the syntactic function of the NP (subject, object, PP adjunct, other), the sortal class of the discourse entity to which an NP refers, discourse structure, or topicality of the discourse entities. In this paper, we focus on the first three factors, agreement (Agree), syntactic function (Syn), and sortal class (Class).

Since we are using syntactically annotated data in the Penn Treebank-II format, the syntactic function of an NP was derived from these annotations. Agreement for gender, number, and person was labelled by hand. Since English has almost no nominal case morphemes, case was not annotated.

Sortal classes provide information about the discourse entity that a referring expression evokes or accesses. The classes, summarized in Table 2, were derived from EuroWordNet BaseTypes (Vossen, 1998) and are defined extensionally on the basis of WordNet synsets. Their selection was motivated by two main considerations: all classes should occur in all genres, and the number of classes should be as small as possible in order to avoid problems with sparse data. Four classes, State, Event, Action, and Property, cover different types of situations, two cover spatiotemporal characteristics of situations (Loc/Time). The four remaining classes cover the two dimensions "concrete vs. abstract (Concept)" and "human (Pers) vs. non-human (PhysObj) vs. institutionalised groups of humans (Group)".

Since we are only interested in the decision whether to employ pronouns rather than full NPs and less in the form of the NP itself, and since our methodology is based on corpus annotation, we did not take into account more formal semantic categories such as kinds vs. individuals.

Co-specification-level factors depend on information about sequences of referring expressions

Person	one or more human beings
Group	institutionalized group of human beings
PhysObj	physical object
Concept	abstract concept
Loc	geographical location
Time	date, time span
Event	sth. which takes place in space and time
Action	sth. which is done
State	state of affairs, feeling, ...
Property	characteristic or attribute of sth.

Table 2: Overview of Sortal Classes with rough characterizations of relevant synsets

which co-specify with each other. Such a sequence consists of all referring expressions that evoke or access the same discourse entity. In this paper, we use the following factors from the literature: distance to last mention (Dist and Dist4), ambiguity (Ambig), parallelism (Par), form of the antecedent (FormAnte), and syntactic function of the antecedent (SynAnte). We also distinguish between discourse entities that are only evoked once, *deadend* entities, and entities that are accessed repeatedly.

Parallelism is defined on the basis of syntactic function: a referring expression and its antecedent are parallel if they have the same syntactic function.

For calculating distance and ambiguity, we segmented the texts into major clause units (MCUs). Each MCU consists of a major clause C plus any subordinate clauses and any coordinated major clauses whose subject is the same as that of C and where that subject has been elided.

Dist provides the number of MCUs between the current and the last previous mention of a discourse entity. When an entity is evoked for the first time, Dist is set to "D". Dist4 is derived from Dist by assigning the fixed distance 2 to all referring expressions whose antecedent is more than 1 MCU away. Ambiguity is defined as the number of all discourse entities with the same agreement features that occur in the previous unit or in the same unit before the current referring expression.

3 Data

Our data consisted of twelve (plus two) texts from the Brown corpus and the corresponding part-of-speech and syntactic annotations from the Penn Treebank (LDC, 1995). The texts were selected because they contained relatively little or no direct speech; segments of direct speech pose problems for both pronoun resolution and generation because of the change in point of view. Morpho-syntactic information such as markables, part-of-speech labels, grammatical role labels, and form of referring expression were automatically extracted from the existing Treebank annotations.

The texts come from four different genres: Popular Lore (CF), Belles Lettres (CG), Fiction/General (CK), and Fiction/Mystery (CL). The choice of genres was dictated by the availability of detailed Treebank-II parses. Table 3 shows that the distribution of referring expressions differs considerably between genres.

The texts from the two non-narrative types, CF and CG, contain far more discourse entities and far less pronouns than the narrative genres CK and CL. The high number of pronouns in CK and CL is partly due to the fact that in one text from each genre, we have a first person singular narrator. CK patterns with CF and CG in the average number of MCUs; the sentences in the sample from mystery fiction are shorter and arguably less complex. CL also has disproportionally few deadend referents. The high percentage of deadend referents in CK is due to the fact that two of the texts deal with relationship between two people. These four discourse referents account for the 4 longest coreference chains in CK (85, 96, 109, and 127 mentions).

Two annotators (the authors, both trained linguists), hand-labeled the texts with co-specification information based on the specifications for the Message Understanding Coreference task (Hirschman & Chinchor (1997); for theoretical reasons, we did not mark reflexive pronouns and appositives as co-specifying). The MCUs were labelled by the second author. All referring expressions were annotated with agreement and sortal class information. Labels were placed using the GUI-based annotation tool REFEREE (DeCristofaro et al., 1999).

The annotators developed the Sortal Class annotation guidelines on the basis of two training texts. Then, both labellers annotated two texts from each genre independently (eight in total). These eight texts were used to determine the reliability of the sortal class coding scheme. Since sortal class annotation is intrinsically hard, the annotators looked up the senses of the head noun of each referring NP that was not a pronoun or a proper name in WordNet. Each sense was mapped directly to one or more of the ten classes given in Table 2. The annotators then chose the adequate sense.

The reliability of the annotations were measured

Genre	words	ref. expr.	entities	sequ.	MCUs	% pron.	% deadend	med. len.
CF	6097	1725	1223	125	304	19.59% (1.8%, 0.3%, 58.3%)	89.78%	3
CG	6103	1707	1290	120	269	16.17% (9.8%, 1.1%, 4%)	90.70%	2
CK	6020	1848	1071	113	386	36.15% (19.5%, 1.2%, 56.1%)	89.45%	2
CL	6018	1846	954	170	477	35.64% (14.0%, 1.5%, 53.6%)	80.09%	4

Table 3: Relevant quantitative characteristics of the texts. Average length: 2020 words, 120 MCUs. *sequ.:* number of sequences of co-specifying referring expressions. *% deadend:* percentage of discourse entities mentioned only once. *% pronouns:* percentage of all referring expressions realized as pronouns, in brackets: perc. of first person singular pronouns, perc. of second person singular pronouns, perc. of third person singular masculine and feminine pronouns. *med. len.:* median length of sequences of co-specifying referring expressions

with Cohen's κ (Cohen, 1960; Carletta, 1996). Cohen (1960) shows that a κ between 0.68 and 0.80 allows tentative conclusions, while $\kappa > 0.80$ indicates reliable annotations. For genres CF ($\kappa = 0.83$), CK ($\kappa = 0.84$) and CL ($\kappa = 0.83$), the sortal class annotations were indeed reliable, but not for genre CG ($\kappa = 0.63$). Nevertheless, overall, the sortal class annotations were reliable ($\kappa = 0.8$). Problems are mainly due to the abstract classes Concept, Action, Event, State, and Property. Abstract head nouns sometimes have several senses that fit the context almost equally well, but that lead to different sortal classes. Another problem is metaphorical usage. This explains the bad results for CG, which features many abstract discourse entities.

4 Towards a Probabilistic Genre-Independent Model

In this section, we investigate to what extent the factors proposed in section 2.2 influence the decision to prominalize. For the purpose of the statistical analysis, pronominalization is modelled by a feature Pro. For a given referring expression, that feature has the value "P" if the referring expression is a personal or a possessive pronoun, else "N". We model this variable with a binomial distribution. [1]

4.1 How do the Factors Affect Pronominalization?

First, we examine for all nine factors if there is a statistical association between these factors and Pro. Standard non-parametric tests show a strong association between all nine factors and Pro.[2] This holds both for all referring expressions and for those that occur in sequences of co-specifying referring expressions. All of the tests were significant at the $p < 0.001$-level, with the exception of Par: for expressions that are part of co-specification sequences the effect of that factor is not significant.

In the next analysis step, we determine which of the feature values are associated disproportionally often with pronouns, and which values tend to be associated with full NPs. More specifically, we test for each feature-value pair if the pronominalization probability is significantly higher or lower than that computed over (a) the complete data set, (b) all referring expressions in sequences of co-specifying referring expressions, (c) all third person referring expressions in sequences. Almost all feature values show highly significant effects for (a) and (b), but some of these effects vanish in condition (c). Below, we report on associations which are significant at $p < 0.001$ under all three conditions.

Unsurprisingly, there is a strong effect of agreement values: NPs referring to the first and second person are always pronominalized, and third person masculine or feminine NPs, which can refer to persons, are pronominalized more frequently than third person neuter and third person plural. Pronouns are strongly preferred if the distance to the antecedent is 0 or 1 MCUs. Referring expressions are more likely to be pronominalized in subject position than as a PP adjunct, and referring expressions with adjuncts as antecedents are also pronominalized less often than those with antecedents in subject or object position. There is a clear preference for pronouns as possessive determiners, and referring expressions that co-specify with an antecedent possessive pronoun are highly likely to be pronominalised. We

[1] For all statistical calculations and for the logistic regression analyses reported below, we used R (Ihaka & Gentleman, 1996).

[2] We used the Kruskal-Wallis test for the ordinal Ambig variable and the χ^2-test for the other, nominal, variables. Since first mentions and deadends are coded by the character "D" in both Dist and Dist4, both are treated as a categorical variable by R. For more on these tests, see (Agresti, 1990).

also notice strong genre-independent effects of parallelism. Although at first glance, Ambig appears to have a significant effect as well, (median ambiguity for nouns is 3, median ambiguity for pronouns 0), closer inspection reveals that this is mainly due to first and second person and third person masculine and feminine pronouns.

The sortal classes show a number of interesting patterns (cf. Table 4). Not only do the classes differ in the percentage of deadend entities, there are also marked differences in pronominalizability. There appear to be three groups of sortal classes: Person/Group, with the lowest rate of deadend entities and the highest percentage of pronouns – not only due to the first and second person personal pronouns – , Location/PhysObj, with roughly two thirds of all entities not in sequences and a significantly lower pronominalization rate, and Concept/Action/Event/Property/State/Concept, with over 80% deadend entities. Within this group, Action, Event, and Concept are pronominalized more frequently than State and Property. Time is the least frequently pronominalized class. An important reason for the difference between Loc and Time might be that Times are almost always referred back to by temporal adverbs, while locations, especially towns and countries, can also be accessed via third person neuter personal pronouns.

Interactions between the factors and genre were examined by an analysis of deviance run on a fitted logistic regression model; significance was calculated using the F-test. All factors except for Par show strong ($p < 0.001$) interactions with Genre. In other words, the influence of all factors but parallelism on pronominalization is mediated by Genre. There are two main reasons for this effect: first, some genres contain far more first and second person personal pronouns, which adds to the weight of Agree, and second, texts which are about persons and the actions of persons, such as the texts in CK and CL, tend to use more pronouns than texts which are mainly argumentative or expository.

4.2 Which Factors are Important?

To separate the important from the unimportant factors, many researchers use decision and regression trees, mostly the binary CART variant (Breiman et al., 1984). We use a different kind of model here, logistic regression, which is especially well suited for categorical data analysis (cf. eg. Agresti (1990) or Kessler et al. (1997)). In this model, the value of the binary target variable is predicted by a linear combination of the predictor variables. Variable weights indicate the importance of a variable for classification: the higher the absolute value of the weight, the more important it is.

Logistic regression models are not only evaluated by their performance on training and test data. We could easily construct a perfect model of any training data set with n variables, where n is the size of the data set. But we need models that are small, yet predict the target values well. A suitable criterion is the Akaike Information Criterion (AIC, Akaike (1974)), which punishes both models that do not fit the data well and models that have too many parameters. The quality of a factor is judged by the amount of variation in the target variable that it explains. Note that increased prediction accuracy does not necessarily mean an increase in the amount of variation explained. As the model itself is a *continuous* approximation of the *categorical* distinctions to be modelled, it may occur that the numerical variation in the predictions decreases, but that this decrease is lost when re-translating numerical predictions into categorical ones.

The factors for our model were selected based on the following procedure: We start with a model that always predicts the most frequent class. We then determine which factor provides the greatest reduction in the AIC, add that factor to the model and retrain. This step is repeated until all factors have been used or adding another factor does not yield any significant improvements anymore.[3]

This procedure invariably yields the sequence Dist4, Agree, Class, FormAnte, Syn, SynAnte, Ambig, Par, both when training models on the complete data set and when training on a single genre. Inspection of the AIC values suggests that parallelism is the least important factor, and does not improve the AIC significantly. Therefore, we will discard it from the outset. All other factors are maintained in the initial full model. This model is purely additive; it does not include interactions between factors. This approach allows us to filter out factors which only mediate the influence of other factors, but do not exert any significant influence of their own. Note that this probabilistic model only provides a numerical description of how its factors affect pronominalization in our corpus. As such, it is not equivalent to a theoretical model, but rather provides data for fur-

[3]We excluded Dist from this stepwise procedure, since the relevant information is covered already by Dist4, which furthermore has much fewer values.

Class	Act	Concept	Event	Group	Loc	Pers	PhysObj	Prop	State	Time
% deadend	84.1	80.0	88.0	46.1	63.3	17.3	65.5	88.5	87.8	92.9
% pronouns	6.2	8,5	6.0	28.4	5.7	63.4	10.2	2.5	3.2	0.3
% pron. (sequences)	32.5	29.6	33.3	51.6	15.4	73.8	27.2	21.4	23,7	4.5

Table 4: Results for Sortal Classes. % deadend: percentage of deadend entities; % pronouns: percent pronominalised, % pron. (sequences: percent pronominalised relative to all occurrences in co-specification sequences

	CF	CG	CK	CL	all
% correct	**97.1**	93.5	93.6	91.5	93.1
AIC	**324.7**	654.8	786.1	904.0	2685.8
% variation	**83.0**	65.4	70.1	65.4	68.7

Table 5: Quality of models fitted to each of the genre-specific corpora (CF, CG, CK, CL) and the complete data set (all). % correct: correctly predicted pronominalization decition, AIC: Akaike Information Criterion, % variation: percentage of original variation in the data (as measured by deviance) accounted for by the model

ther theoretical interpretation.

Results of a first evaluation of the full model are summarized in Table 5. The model can explain more than two thirds of the variation in the complete data set and can predict pronominalization quite well on the data it was fitted on. The matter becomes more interesting when we examine the genre-specific results. Although overall prediction performance remains stable, the model is obviously suited better to some genres than to others. The best results are obtained on CF, the worst on CL (mystery fiction). In the CL texts, MCUs are short, a third of all referring expressions are pronouns, there is no first person singular narrator, and most paragraphs which mention persons are about the interaction between two persons.

The Relative Importance of Factors. All values of Dist4 have very strong weights in all models; this is clearly the most important factor. The same goes for Agree, where the first and second person are strong signs of pronominalization, and, to a lesser degree, masculine and feminine third person singular. The most important distinction provided by Class appears to be that between Persons, non-Persons, and Times. This holds as well when the model is only trained on third person referring expressions. For singular referring expressions, Personhood information is reflected in gender, but not for plural referring expressions. Another important influence is the form of the antecedent. The syntactic function of the referring expression and of its antecedent are less important, as is ambiguity.

In order to examine the importance of the factors in more detail, we refitted the models on the complete data set while omitting one or more of the three central features Dist4, Agree, and Class. The results are summarized in Table 6. The most interesting finding is that even if we exclude all three factors, prediction accuracy only drops by 3.2%. This means that the remaining 4 factors also contain most of the relevant information, but that this information is coded more "efficiently", so to speak, in the first three. Speaking of these factors, questions concerning the effect of sortal class remains. Remarkably enough, when sortal class is omitted, accuracy *increases* by 0.7%. The increase in AIC can be explained by a decrease in the amount of explained variation. A third result is that information about the *form of the antecedent* can substitute for distance information, if that information is missing. Both variables code the crucial distinctions between expressions that evoke entities and those that access evoked entities. Furthermore, a pronominal antecedent tends to occur at a distance of less than 2 MCUs. The contribution of syntactic function remains stable and significant, albeit comparatively unimportant.

Predictive Power: To evaluate the predictive power of the models computed so far, we determine the percentage of correctly predicted pronouns and NPs. The performance of the trained models was compared to two very simple algorithms:

Algorithm A: Always choose the most frequent option (i.e. noun).

Algorithm B: If the antecedent is in the same MCU, or if it is in the previous MCU and there is no ambiguity, choose a pronoun; else choose a noun.

Table 7 summarises the results of the comparison. To determine the overall predictive power of

excluded	fit		% explained variation						
	AIC	%correct	Dist4	Agree	Class	PForm	Syn	PSyn	Ambig
none	2686	92.6	54.4	21.1	5.7	3.8	2.3	*0.5*	1.1
Class	2785	**93.3**	54.4	21.1	n.a.	4.7	2.8	0.5	1.1
Agree	2984	**92.6**	54.4	n.a.	**14.3**	6.2	2.7	0.6	1.1
Dist4	3346	90.2	n.a.	35.8	6.1	**32**	3	0.8	*0.1*
Dist4 + Class	3443	90.2	n.a.	35.8	n.a.	**33.7**	3.4	0.8	*0.1*
Dist4 + Agree	3597	89.6	n.a.	n.a.	**31.4**	**35.4**	3.1	0.8	*0.2*
Agree + Class	3098	**92.6**	54.4	n.a.	n.a.	13.11	3.5	*0.5*	3.6
Dist4 + Agree + Class	3739	89.4	n.a.	n.a.	n.a.	**52.62**	4	0.7	1.7

Table 6: Effect of leaving out any one of the three most important factors on model fit. *italics:* significance is $p < 0.05$, for all other factors, $p < 0.005$ or better.

	test data set				
	CF	CG	CK	CL	all
Alg. A	80.4	83.8	63.8	65.4	72.8
Alg. B	91.1	**93.0**	88.6	84.7	89.4
Model	96.5	92.2	**91.8**	**90.9**	92.6 ± 0.02
w/o Class	**96.8**	92.4	91.7	90.7	$\mathbf{93.0 \pm 0.01}$

Table 7: Results of algorithms vs. models on test data in % correct prediction if referring expression is to be pronominalised or not. Setup for genres: model is trained on three genres, tested on the remaining one

the model, we used 10-fold cross-validation. Algorithm A always fares worst, while algorithm B, which is based mainly on distance, the strongest factor in the model, performs quite well. Its overall performance is 3.2% below that of the full model, and 3.6% below that of the full model without sortal class information. It even outperforms the models on CG, which has the lowest percentage of Persons (12.9% vs. 35% for CF and 43.4% and 43.5% for CL and CK). For all other genres, the statistical models outperform the simple heuristics. Excluding sortal class information can boost prediction performance on unseen data by as much as 0.4% for the complete corpus. The apparent contradiction between this finding and the results reported in the previous section can be explained if we consider that not only were some sortal classes comparatively rare in the data (Property, Event), but that our sortal class definition may still be too fine-grained.

We evaluated the genre-independence of the model by training on three genres and testing on the fourth. The results show that the model fares quite well for genre CF, which is also the genre where the overall fit was best (see Table 5). We therefore hypothesize that the decrease in performance is mainly due to the model itself, not to the training data. The results presented in both Table 5 and 7 show that although the model we have found is not quite as genre-independent as we would want it to be, it provides a reasonable fit to all the genres we examined.

5 Future Work

We have described a probabilistic model of pronominalization that is able to correctly predict 93% of all pronouns in a corpus that consists of twelve texts from four different genres. Since the model was derived from a limited corpus and a limited number of genres, we cannot guarantee that our results are applicable to all texts without modifications. But since its performance on our sample is consistently above 90% correct, we are reasonably confident that our main findings will hold for a wide variety of texts and text types. In particular, we isolated several factors which are robust predictors of pronominalization across genres: distance from last mention and agreement, and to a certain extent the form of the antecedent, which appears to be a good substitute if the other two factors are not available. All three features can be computed on the basis of a chunk parse, a rough morphosyntactic analysis of the resulting NPs, and co-specification sequences. In computational terms, they are comparatively cheap. Large corpora can be annotated relatively quickly with this information, which can then be used for statistical pronoun generation.

The comparatively expensive sortal class annotation, on the other hand, was not very important in the final model; in fact, prediction accuracy decreased when sortal class was included. There are two main reasons for this: first, the proposed sortal class annotation scheme needs further work,

second, the relationship between sortal class and pronominalization may well be too intricate to be modelled by the factor Class alone.

We set out to find a genre-independent model of pronominalization. The model we found performs quite well, but genre still considerably affects its performance. Where does the remaining, unexplained variation come from? The variation might be just that – stylistic variation. It might stem from one of the traditional factors that we did not take into account here, such as thematic role. However, we suspect that the crucial factor at play here is discourse structure (McCoy & Strube, 1999).

Acknowledgements Work on this paper was begun while Michael Strube was a postdoctoral fellow at the Institute for Research in Cognitive Science, University of Pennsylvania, and Maria Wolters visited the Institute for a week in summer 1999. We would like to thank Kathleen McCoy, Jonathan DeCristofaro, and the three anonymous reviewers for their comments on earlier stages of this work.

References

Agresti, Alan (1990). *Categorical Data Analysis*. New York, N.Y.: Wiley.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions Automatic Control*, 19:716–722.

Azzam, Saliha, Kevin Humphreys & Robert Gaizauskas (1998). Evaluating a focus-based approach to anaphora resolution. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* Montréal, Québec, Canada, 10–14 August 1998, pp. 74–78.

Breiman, Leo, Jerome H. Friedman, Charles J. Stone & R.A. Olshen (1984). *Classification and Regression Trees*. Belmont, Cal.: Wadsworth and Brooks/Cole.

Cardie, Claire & Kiri Wagstaff (1999). Noun phrase coreference as clustering. In *Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora,* College Park, Md., 21–22 June 1999, pp. 82–89.

Carletta, Jean (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Cohen, Jacob (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Dale, Robert (1992). *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. Cambridge, Mass.: MIT Press.

DeCristofaro, Jonathan, Michael Strube & Kathleen F. McCoy (1999). Building a tool for annotating reference in discourse. In *ACL '99 Workshop on the Relationship between Discourse/Dialogue Structure and Reference, University of Maryland, Maryland, 21 June, 1999*, pp. 54–62.

Fellbaum, Christiane (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Fraurud, Kari (1996). Cognitive ontology and NP form. In T. Fretheim & J. Gundel (Eds.), *Reference and Referent Accessibility*, pp. 65–87. Amsterdam, The Netherlands: Benjamins.

Ge, Niyu, John Hale & Eugene Charniak (1998). A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora,* Montréal, Canada, pp. 161–170.

Grosz, Barbara J., Aravind K. Joshi & Scott Weinstein (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Hirschman, Lynette & Nancy Chinchor (1997). *MUC-7 Coreference Task Definition,* `http://www.muc.sais.com/proceedings/`.

Ihaka, Ross & Ross Gentleman (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Kessler, Brett, Geoffrey Nunberg & Hinrich Schütze (1997). Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics,* Madrid, Spain, 7–12 July 1997, pp. 32–38.

LDC (1995). *Penn Treebank-II*. Linguistic Data Consortium. University of Pennsylvania, Philadelphia, Penn.

McCoy, Kathleen F. & Michael Strube (1999). Generating anaphoric expressions: Pronoun or definite description? In *ACL '99 Workshop on the Relationship between Discourse/Dialogue Structure and Reference, University of Maryland, Maryland, 21 June, 1999*, pp. 63–71.

Poesio, Massimo, Renate Henschel, Janet Hitzeman & Rodger Kibble (1999). Statistical NP generation: A first report. In R. Kibble & K. van Deemter (Eds.), *Proceedings of the Workshop on The Generation of Nominal Expressions, 11th European Summer School on Logic, Language, and Information, Utrecht, 9-13 August 1999*.

Sidner, Candace L. (1983). Focusing in the comprehension of definite anaphora. In M. Brady & R.C. Berwick (Eds.), *Computational Models of Discourse*, pp. 267–330. Cambridge, Mass.: MIT Press.

Vossen, Piek (Ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht, The Netherlands: Kluwer.

Yeh, Ching-Long & Chris Mellish (1997). An empirical study on the generation of anaphora in Chinese. *Computational Linguistics*, 23(1):169–190.